VALIDATION AND IMPROVEMENT OF AUTOMATIC PHONETIC TRANSCRIPTIONS

Catia Cucchiarini & Diana Binnenpoorte

Department of Language and Speech,
University of Nijmegen, The Netherlands
{c.cucchiarini, d.binnenpoorte}@let.kun.nl

ABSTRACT

The ultimate aim of our research is to show that good-quality phonetic transcriptions of large speech corpora can be obtained by employing automatic techniques initially developed for ASR. The experiment presented in this paper has two aims. The first is to show how the quality of an automatic transcription that is easily obtained through lexicon lookup can be measured in a way that is methodologically sound. The second is to show how, while measuring the quality of an automatic transcription, it is possible to obtain information that can subsequently be used to improve the automatic transcription where necessary. As a result, correction by human transcribers should become more efficient or even superfluous.

1. INTRODUCTION

Large speech corpora are becoming available for research and applications. Before these data can be used for their purpose though, they often need to be annotated. Manual annotation of such large speech corpora is time-consuming and costly, and for this reason often impossible. Recourse to automatic or semi-automatic annotation is therefore desirable. In the case of phonetic transcriptions, manual annotation is particularly time-consuming so that automatic annotation, if of reasonable quality, would certainly be preferred. However, the point is exactly how to establish that a certain type of automatic annotation is of acceptable quality.

The only way seems to be to compare automatic annotations with annotations produced by human beings. However, in the specific case of phonetic transcription this is not straightforward, because phonetic transcriptions made by human transcribers tend to contain an element of subjectivity [1]. In other words, to establish whether an automatically generated phonetic transcription is of satisfactory quality, it is necessary to do something more than just comparing this transcription to one produced by a human transcriber.

The ultimate aim of our study is to show that good-quality phonetic transcriptions of large corpora can be obtained with greater efficiency by employing automatic techniques that have been initially developed for automatic speech recognition (ASR). Although we do not believe that state-of-the art technology allows completely automatic phonetic annotation for all speech styles, we are nevertheless convinced that resorting to semi-automatic annotation can improve efficiency without penalizing quality. At the same time, since we are aware of the difficulties involved in assessing automatic transcription quality, we would like to propose a procedure that could be used for this purpose. To pursue our objectives, in this paper we present an investigation that has two aims. The first is to show how the quality of an easily obtained (through lexicon look-up) automatic transcription can be measured in a way that is methodologically sound. The second aim is to show how, while measuring the quality of an automatic transcription, it is possible to obtain information that can subsequently be used to improve the automatic transcription where necessary. The improvement procedure we propose is first of all automatic because this is more efficient. It is based on pronunciation variation modeling. For certain speech styles it may turn out that it is not possible to further improve transcription quality automatically and that correction by human transcribers is required. However, starting from an optimal automatic transcription the transcribers will have to correct fewer symbols and the whole procedure will be less costly.

To summarize, in the rest of this paper we propose a procedure for transcription validation and improvement that will have the effect of making human intervention more efficient or even superfluous, so that good-quality phonetic transcriptions of large amounts of material can be obtained at lower costs.

2. MEASURING THE QUALITY OF AUTOMATIC TRANSCRIPTIONS

Since in large speech corpora phonetic transcriptions constitute the basis for further processing (research, ASR training, etc.), they can be viewed as representations or measurements of the speech signal and it is therefore legitimate to ask to what extent they live up to the quality standards of reliability and validity that are required of any form of measurement. With respect to automatic transcriptions, the problem of quality assessment is complex because comparison with human performance, which is customary in many fields, is not straightforward owing to the subjectivity of human transcriptions and to a series of methodologically complex issues that will be explained below.

2.1. Reliability and validity of human and automatic phonetic transcriptions

In general terms, the reliability of a measuring instrument represents the degree of consistency observed between repeated measurements of the same object made with that instrument. It is an indication of the degree of accuracy of a measuring device. Validity, on the other hand, is concerned with whether the instrument measures what it purports to measure. In fact, the definitions of reliability and validity used in test theory are much more complex and will not be treated in this paper. The description provided above indicates an important difference between the reliability of human as opposed to automatic transcriptions and is related to the fact that human

transcriptions suffer from intra-subject variation and repeated measurements of the same object will differ from each other. With automatic transcriptions this can be prevented because a machine can be programmed in such a way that repeated measurements of the same object always give the same result, thus yielding a reliability coefficient of 1, the highest possible. It follows that with respect to the quality of automatic transcription only one, albeit not trivial, question needs to be answered, viz. that concerning validity. The description of validity given above suggests that any validation activity implies the existence of a correct representation of what is to be measured, a so-called benchmark or 'true' criterion score (as in test theory). Since such a 'true' transcription cannot be obtained, because human transcriptions are subject to error, one can at best try to approach the ideal reference. For instance, one cannot establish the validity of an automatic transcription simply by comparing it with an arbitrarily chosen human transcription, because the latter would inevitably contain errors. Unfortunately, this seems to be the normal practice in studies on automatic transcription. Furthermore, studies in which multiple human transcribers were involved seem to suffer from other methodological limitations. For instance, in [2] an attempt was made at validating an automatic transcription by comparing this with transcriptions of the same material made by three different transcribers. The three human transcriptions were first compared to each other with a view to obtaining an upper bound for automatic transcription quality: the degree of agreement observed between human transcribers would represent the highest attainable between human and automatic transcription. As was expected, the degree of agreement in the comparisons between the three human transcriptions, on average 94.8%, appeared to be higher than that in comparisons between automatic and human transcriptions, on average 88.4%. However, the authors fail to relate this finding to the fact that the human transcribers in their experiment were given the citation forms as a guide to their transcriptions. This introduces a bias in the procedure which has the effect of inflating the degree of agreement between human transcriptions. Although this constitutes a weak point in their validation procedure, it also implies that the performance of their system is probably better than they maintain in their paper: they just overestimated the performance of the human transcribers. However, this also means that the human agreement values they report are not representative.

A viable solution to the transcription validation problem has been proposed by [1] who suggest using a consensus transcription as the point of reference. A consensus transcription is a transcription made by at least two experienced phoneticians after having reached a consensus on each symbol contained in the transcript. The fact that different transcribers are involved and that they have to reach a consensus before writing down the symbols may be seen as an attempt to minimise errors of measurement, thus approaching 'true' criterion scores.

2.2. A procedure for validating and improving automatic phonetic transcriptions

In this section we suggest a procedure for validating and further improving automatic phonetic transcriptions that stems from the operationalisation of transcription validity presented above and is based on the assumption that the best benchmark for an automatic transcription is a consensus transcription made by two or more experienced transcribers. In this procedure the consensus transcription will serve as the reference transcription (T*ref*) to which the automatic transcription (T*aut*) is compared to determine how much they differ from each other. The degree of agreement observed between T*aut* and T*ref* then has to be related to the degree of agreement that is generally observed between human transcriptions that are of the same level of detail and that are not made under biased conditions, because this agreement level constitutes the upper bound, as in the study reported in [2]. If the degree of agreement between T*aut* and T*ref* is higher than what usually observed between human transcriptions, one could accept T*aut* as is; alternatively, if the degree of agreement between T*aut* and T*ref* is lower than what usually observed between human transcriptions, one could first try to improve T*aut* so as to make it more similar to T*ref* Therefore the comparison between T*aut* and T*ref* should be made in such a way that information can be obtained not only on the number, but also on the nature and frequency of the discrepancies. To this end the ALIGN program [3] is used in our procedure. Furthermore, information on the frequency of occurrence of various phonological processes in T*ref* should be obtained to establish how these processes can best be represented in T*aut*. In this respect we distinguish between static and dynamic modeling. In static modeling only the most frequent variant is included in the lexicon. This is likely to improve the quality of T*aut* when the relative frequency of occurrence (F*rel*) of such binary processes is either very low or very high [4]. F*rel* is calculated by dividing the number of times a process is applied by the number of times the process could have been applied because the conditions for application were met. Alternatively, when F*rel* is between 40% and 60 %, static modeling will not work and one has to resort to dynamic modeling, which implies that multiple pronunciation variants are included in the lexicon and that a continuous speech recognizer (CSR) is allowed to determine which one of these variants best matches the speech signal. Finally, in order to establish whether modeling pronunciation variation, either statically or dynamically, has indeed improved the quality of T*aut*, one can again compare T*aut* with T*ref*.

3. EXPERIMENT ON VALIDATION AND IMPROVEMENT OF AUTOMATIC TRANSCRIPTIONS

To test the validation and improvement procedure outlined above, an experiment was carried out which had two goals: 1) determining how much a T*aut* obtained through lexicon-lookup differs from T*ref* (validation) and 2) establishing in what respects T*aut* deviates from T*ref*, so that T*aut* can be optimized until it is good enough so as to make human intervention more efficient or even superfluous (improvement).

3.1. Materials and implementation

3.1.1. Speech material and speakers

The speech material to be transcribed was taken from the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) project, which is aimed at compiling a large (10 million words) corpus of spoken Dutch from the Netherlands and Flanders.

The whole corpus will be orthographically transcribed and annotated at various linguistic levels [5]. The present experiment was limited to the varieties spoken in the Netherlands. The subcorpus was selected so as to obtain variation in speech style and speaker. It contains fragments of four speech styles (read speech (RS), lectures (LC), interviews (IN), and spontaneous conversations (SC)) produced by twenty speakers (eleven males and nine females, age between 20 and 73) from different country regions. In this way a plausible sample of Northern Dutch was obtained (see Table 1).

speech style	mode	# of words	duration
RS	monologue	682	04:57 min
LC	monologue	892	05:09 min
IN	dialogue	523	03:01 min
SC	dialogue	615	03:01 min
Total		2712	16:08 min

Table 1 Overview of the speech material

3.1.2. Tref, Taut and the Alignment

The consensus transcription (T*ref*) was made by two experienced transcribers who transcribed together from scratch without using the orthographic transcription of the material as a guideline. They used the CGN symbol set which is derived from the SAMPA set for Dutch and does not contain diacritics. It took the transcribers about 60 hours to complete the T*ref* for the 16 minutes of speech selected for the experiment.

Since our aim is to show that simple automatic techniques can boost efficiency in producing phonetic transcriptions, we deliberately started with the most simple automatic transcription: one obtained by concatenating the canonical phonetic representations obtained from the CGN lexicon through a lexicon-lookup procedure. The transcriptions in the lexicon were obtained by means of TREETALK [6], a grapheme-to-phoneme converter trained on Celex. In the resulting phonetic representations all so-called obligatory word-internal processes [7] were applied, whereas optional word-internal processes were not applied. With the sole exception of degemination [7], cross-word processes were not applied in this concatenated T*aut*.

T*aut* and T*ref* were subsequently aligned by means of the ALIGN program, which calculates the distance between corresponding phonemes on the basis of articulatory features like place and manner of articulation, voice, lip rounding, length, etc. so that substituting a /t/ for a /d/ has a lower cost than substituting a /t/ for a /x/. The output of the program exactly specifies in what respects T*aut* differs from T*ref*, and for each type of discrepancy, be it a deletion, an insertion or a substitution, it indicates which articulatory features are concerned. In this way it is possible to determine what should be changed in T*aut* to make it more similar to T*ref*

3.2. Results: validation of T*aut*

The frequency of phone substitutions, deletions and insertions as calculated by the ALIGN program is shown in Table 2. To interpret these data we have to compare them to those concerning transcriptions by human transcribres. Data on agreement between human transcribers appear to vary between 93.1% and 94.4% (which correspond to deviation percentages of 6.9% and 5.6%, respectively) for careful speech [8], and between 78.8% and 86.2% (which correspond to deviation percentages of 21.2% and 13.8%, respectively) for less careful speech [9]. The data in Table 2 indicate that the quality of this initial T*aut* is already reasonable, especially if we consider that the agreement data reported in [8 and 9] are probably inflated, as explained above.

category	substitutions	deletions	insertions	Total
RS	6.4 %	1.9 %	4.2 %	12.5 %
LC	8.3 %	3.4 %	7.1 %	18.8 %
IN	7.3 %	4.0 %	8.8 %	20.1 %
SC	9.4 %	2.5 %	12.4 %	24.3 %

Table 2 Deviations per speech style for T*aut*1

3.3. Results: improvement of Taut

From Table 2 it can be inferred that T*aut* could be improved if some of the processes causing discrepancies between T*aut* and T*ref* were modeled. The potential of such an intervention was investigated in a qualitative analysis of the outcome of the ALIGN program and of T*ref* that was presented in [10]. Owing to space limitations, this analysis cannot be presented here so we limit ourselves to the results that are implemented in this paper. The study revealed a high percentage of word-boundary voice substitutions, about half of the total number of substitutions, while calculations of the relative frequency of application (F*rel*) of cross-word voice assimilation processes in T*ref*, indicated that this process is frequently applied in all four speech styles investigated. Such high values of F*rel* suggest that if the other variant (in this case the one with voice assimilation applied) were chosen, T*aut* would approach T*ref* more closely. In [10] we investigated how this form of static modeling would potentially affect the degree of agreement between T*aut* and Tref and found that the percentages of substitutions could be reduced so as to bring the total percentages of deviations to 10.7% (RS), 16.9% (LC), 18.3% (IN), and 21.7% (SC). As mentioned above, the research presented in [10] revealed other possibilities of improving T*aut* which will certainly be implemented in the future. As a first step, however, we decided to focus on static modeling of cross-word voice assimilation since this appeared to have a great potential of improving T*aut* - the reductions mentioned above are indeed substantial - and can be implemented very easily. The application of cross-word voice assimilation produced new degemination contexts so that this process had to be applied a second time. Subsequently, we again aligned the newly obtained T*aut* with T*ref*

%	RS	LC	IN	SC
S T*aut*2	5.7	6.9	5.7	7.8
D T*aut*2	2.1	1.5	1.5	1.8
I T*aut*2	2.5	5.7	6.6	11.3
TOTAL				
Predicted	10.7	16.9	18.3	21.7
T*aut*2	10.3	14.1	13.8	20.9

Table 3 Deviations per speech style for T*aut*2

The results are displayed in Table 3, which shows the percentages of substitutions (S), deletions (D) and insertions (I) per speech style, the total percentages of deviations predicted on the basis of the analysis in [10], and the total percentages of deviations obtained with T*aut*2. As can be observed, modeling cross-word voice assimilation improves the quality of T*aut*2 across the board: the percentages of deviations for T*aut*2 (Table 3) are lower than those for T*aut*1 (Table 2), and are even lower than those predicted on the basis of the qualitative analysis in [10].

The additional reductions are probably due to the fact that the predictions in [10] were based only on the potential reductions in the percentages of substitutions, whereas modeling cross-word voice assimilation and again applying degemination also affected the percentages of deletions and insertions, as is clear from Table 3.

4. GENERAL DISCUSSION

In this paper we have proposed and tested a procedure for validating and improving, when necessary, a phonetic transcription that has been generated automatically. In this experiment we deliberately started with an automatic transcription that can be obtained very easily through a simple lexicon-lookup procedure. Two prerequisites for applying this procedure are 1) that the orthographic transcription of the speech material is available, but this is usually the case in many speech corpora; and 2) that a Tref of a representative subsample of the material be first made.

The validation part of our study has revealed that, for some types of speech, such a simple T*aut* as used in our experiment already achieves reasonable quality levels, because the deviations observed between the T*aut* and a T*ref* of the same material are in the order of magnitude of the deviations found between human transcriptions. In the improvement part of our study we have showed that a first, substantial improvement of such a simple T*aut* can be obtained through an equally simple intervention: static modeling of cross-word voice assimilation (both regressive and progressive). As a matter of fact, this intervention reduces the number of substitutions dramatically, and for some speech styles, i.e. RS, the agreement levels observed between T*aut* and T*ref* after applying cross-word voice assimilation are comparable to the levels of agreement normally observed between human transcriptions. For the other speech styles it is clear that even this improved T*aut* is not good enough and that some other measures should be taken.

Our next suggestion would be to apply other automatic techniques that are slightly more complex than static modeling, but in any case require lesser effort than human correction. One such technique would be dynamic modeling of those processes that cannot be modeled statically, because F*rel* is not in the required range. This entails allowing multiple pronunciation variants in the lexicon and having a CSR in forced recognition mode decide which of the pronunciation variants of one and the same word best fits the acoustic signal [11]. It is possible that for certain speech styles even this kind of modeling will not lead to transcriptions of sufficient quality and that human intervention is inevitable. However, thanks to the automatic improvement procedure correction by human transcribers will be more efficient - because fewer symbols will have to be corrected - and therefore less costly.

5. CONCLUSIONS

The results presented in this paper have revealed that the procedure proposed for validating and, if necessary, improving automatic phonetic transcriptions can indeed successfully be used for these purposes. With respect to our more general objective, that of showing how ASR techniques can contribute to increasing efficiency in producing good-quality phonetic transcriptions, we have certainly demonstrated that it is worthwhile considering the possibility of optimizing automatic transcriptions before embarking on costly enterprises with human transcribers, which do not necessarily guarantee higher quality transcriptions.

6. REFERENCES

[1] Shriberg L. D., and Lof, L. "Reliability studies in broad and narrow phonetic transcription", *Clinical Linguistics and Phonetics*, 5, 225-279, 1991.

[2] Wesenick, M-B., and Kipp, A. "Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals", *Proceedings of ICSLP 1996*, Philadelphia, USA, 129-132, 1996.

[3] Cucchiarini, C. "Assessing transcription agreement: metho-dological aspects", *Clinical Linguistics & Phonetics*, 2, 131-155, 1996.

[4] Kessens, J.M., Cucchiarini, C. and Strik, H. A data-driven method for modeling pronunciation variation. Submitted to Speech Communication, 2002.

[5] Oostdijk, N. "The Spoken Dutch Corpus: Overview and first Evaluation", *Proceedings LREC*, Athens, 887-893, 2000.

[6] Hoste, V., Daelemans, W., Tjong Kim Sang, E. and Gillis, S. "Meta-learning for phonemic annotation of corpora", *Proceedings of ICML-2000*, P. Langley (ed), 375-382. Stanford University, 2000.

[7] Booij, G., The phonology of Dutch, Clarendon Press, Oxford, 1995.

[8] Kipp, A., Wesenick, M-B., and Schiel, F. "Automatic detection and segmentation of pronunciation variants in German speech corpora", *Proceedings ICSLP 1996*, Philadelphia, USA, 106-109, 1996.

[9] Kipp, A., Wesenick, B. and Schiel, F. "Pronunciation modeling applied to automatic segmentation of spontaneous speech", *Proceedings EUROSPEECH '97*, 1023-1026, 1997.

[10] Cucchiarini, C., Binnenpoorte, D. and Goddijn, S. "Phonetic Transcriptions in the Spoken Dutch Corpus: how to Combine Efficiency and Good Transcription Quality", *Proceedings EUROSPEECH '01*, 1679-1682, 2001.

[11] Wester, M., Kessens, J.M., Cucchiarini, C. and Strik, H., Obtaining phonetic transcriptions: a comparison between expert listeners and a conituous speech recognizer, *Language and Speech 44 (3)*, 377-403, 2001.

7. ACKNOWLEDGMENTS

This research was supported by the project "Spoken Dutch Corpus (CGN)", which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government.